**Fase I: modelación con la data completa**

**Introducción**

SURA ha venido implementando estrategias de prevención, alineadas al conocimiento integral del individuo desde los aspectos médicos, clínicos y, en un futuro, apuntando al conocimiento de variables más enfocadas al individuo como las psicológicas, de vivienda, etc.

Dentro del conjunto de estrategias mencionadas, se encuentra el índice de desenlaces. Un desenlace puede considerarse como una patología o enfermedad muy impactante: para el individuo por la complejidad y riesgos que ésta representa, y para la compañía por los costos y la atención que esta requiere. Algunos desenlaces son diabetes mellitus, insuficiencia cardiaca crónica, enfermedad renal crónica, cáncer de mama, cáncer de próstata, entre otros.

El proyecto corresponde a la extracción y procesamiento de variables para, posteriormente, construir un modelo predictivo por cada uno de los desenlaces. Cada modelo se ejecuta sobre toda la población de seguros voluntarios y obligatorios, y que no presente el desenlace. De tal manera que se determine qué tan probable es que se desarrolle la patología en los próximos doce meses. Los analistas de negocio, posteriormente, realizan una gestión sobre la población sana pero, que tienen más probabilidad de desarrollar el desenlace.

Este documento contiene las definiciones para la construcción del índice de cáncer de mama, de parte de los equipos de la Gerencia de Salud y EPS SURA, para la población femenina y mayor de 18 años.

**Análisis de literatura**

El cáncer de mama corresponde al crecimiento anormal de algunas células de la mama. Esto puede dar lugar a la expansión anómala a los ganglios linfáticos y, posteriormente, a otros órganos, ocasionando una metástasis. La producción de progesterona y estrógeno puede favorecer el crecimiento de tumores en la mama.

Según la OMS, una de cada 12 mujeres desarrollarán cáncer de mama en algún momento de su vida. En América Latina, entre el 30% y el 40% de casos se detectan en etapa metastásica. En Colombia, entre 2009 y 2019, la mortalidad por esta enfermedad incrementó un 36.5%. La combinación de estrategias de autoexamen, detección temprana y gestión integral del cáncer apuntan a la disminución de la mortalidad y un tratamiento efectivo.

Existen diferentes factores de riesgo que pueden incrementar la probabilidad de, en algún momento de la vida, desarrollar cáncer de mama.

**Datos sociodemográficos**

**Edad:** los estudios muestran que, conforme incrementa la edad edad, se incrementan las probabilidades de alguna mutación anómala en el crecimiento celular. Según la literatura, la mayoría de cánceres de mama se diagnostican después de los 50 años de edad.

**Raza:** la evidencia muestra que las mujeres de raza blanca son más propensas a desarrollar cáncer de mama.

**Hábitos**

**Cigarrillo:** se ha demostrado que algunos componentes del cigarrillo son altamente cancerígenos.

**Alcohol:** el alcohol puede dañar células y hacerlas más propensas al crecimiento anómalo, igualmente, las calorías adicionales que lo componen afectan directamente el peso, el cual está relacionado con la generación de cánceres. Las personas obesas están más propensas a padecerlos.

**Ejercicio:** el ejercicio físico está relacionado inversamente con la obesidad. Las personas obesas están más propensas a padecer distintos tipos de cáncer.

**Resultados clínicos:**

**Índice de masa corporal:** la evidencia muestra que los diferentes tipos de cáncer son más frecuentes en los pacientes obesos. Para ello conviene utilizar el índice de masa corporal, ya que permite cuantificar el peso respecto a la estatura del paciente. Con base en estándares definidos se determina que un paciente tiene sobrepeso cuando su IMC es superior a 25 y obesidad, cuando es superior a 29.

**Antecedentes familiares:**

**Antecedente familiar de cáncer de mama:**

El riesgo de padecer cáncer de mama incrementa si se tienen antecedentes familiares de esta patología tanto de primer como de segundo grado.

**Antecedente familiar de cáncer de cérvix:**

El riesgo de padecer cáncer de mama puede verse influenciado si se tienen antecedentes familiares de cáncer de cérvix o cuello uterino.

**Antecedentes personales:**

**Edad de la menarca:**

Una edad temprana de la menarca (primera menstruación) implica un mayor periodo donde la mujer produce hormonas, lo cual incrementa el riesgo de padecer el cáncer.

**Edad de la menopausia:**

Igualmente, una edad tardía de la menopausia implica un mayor periodo donde la mujer se encuentra produciendo hormonas, lo que incrementa el riesgo del cáncer.

**Edad del primer parto:**

Una edad tardía del primer parto está ligada a una mayor liberación de estrógeno, lo que incrementa el riesgo.

**Embarazo:**

Las mujeres que culminaron satisfactoriamente por lo menos un embarazo tienen menor probabilidad de desarrollar el cáncer.

**Antecedentes clínicos:**

**Menopausia quirúrgica:**

El proceso, conocido como ooforectomía bilateral, implica la remoción de ambos ovarios, lo cual puede favorecer el desarrollo del cáncer.

**Antecedente de biopsia con resultado de hiperplasia ductal:**

Una hiperplasia ductal consiste en el crecimiento anómalo de tejido no cancerígeno en los conductos de la mama. Mujeres que presentan este fenómeno pueden presentar con más probabilidad la aparición del cáncer.

**Mutación de los genes BRC1 y BRC2:**

Estos genes producen proteínas que reparan la información del ADN. La población que posee mutaciones en los genes tiene mucha probabilidad de desarrollar cáncer de mama.

**Terapia de reemplazo hormonal:**

Este procedimiento se aplica a mujeres que están experimentando la menopausia con el fin de alivianar síntomas. En el proceso la mujer consume medicamentos conocidos como estrógenos conjugados. La carga hormonal que se induce puede incrementar el riesgo de padecer el cáncer.

**BIRADS:**

La escala BIRADS corresponde a una clasificación de la severidad que pueden representar las características identificadas en una mamografía. Es una variable categórica que va de 1 a 6, siendo 1 una mamografía sin características raras y 6 un indicio muy sospechoso de una condición cancerígena. Para efectos del análisis se determinó considerar como variables de entrada los índices 4, 4A, 4B, 4C y 5 de la escala.

**Biopsia de mama:**

El procedimiento de biopsia de mama se realiza de acuerdo a indicios que puedan generar sospecha como, por ejemplo, valores altos en la escala BIRADS o antecedentes familiares.

**Valores hormonales:**

La concentración de progesterona, prolactina, estradiol, lutenizante (LH) y folicular (FSH) pueden determinar una probabilidad alta en la aparición del cáncer.

**Definición de variables**

Con base en la revisión de literatura y las herramientas disponibles, se definió el siguiente set de variables:

|  |  |  |
| --- | --- | --- |
| Nombre variable | Grupo Variable | Tipo Variable |
| Asegurado ID | Datos de Identificación | Numérica Discreta |
| Afiliado ID | Datos de Identificación | Numérica Discreta |
| DNI Cliente | Datos de Identificación | Texto |
| Indicador de Negocio | Datos de Identificación | Categórica |
|  |  |  |
| Edad | Sociodemográficas | Numérica Discreta |
| Raza | Sociodemográficas | Categórica |
|  |  |  |
| Cigarrillo | Hábitos | Categórica |
| Alcohol | Hábitos | Categórica |
| Ejercicio | Hábitos | Categórica |
|  |  |  |
| IMC | Resultados clínicos | Numérica Continua |
| Obesidad | Resultados clínicos | Categórica |
| Sobrepeso | Resultados clínicos | Categórica |
| Valor Progesterona, Estradiol, LH,FSH, Prolactina | Resultados clínicos | Numérica |
|  |  |  |
| Hiperplasia Ductal | Antecedentes Personales | Categórica |
| Ooforectomía Bilateral | Antecedentes Personales | Categórica |
| Gen BRC | Antecedentes Personales | Categórica |
| Terapia Hormonal | Antecedentes Personales | Categórica |
| Edad Menopausia | Antecedentes Personales | Numérica Discreta |
| Edad Menarca | Antecedentes Personales | Numérica Discreta |
| Edad Primer parto | Antecedentes Personales | Numérica Discreta |
| Indicador Embarazo | Antecedentes Personales | Categórica |
| Indicador Birads | Antecedentes Personales | Categórica |
| Indicador Biopsia de mama | Antecedentes Personales | Categórica |
|  |  |  |
| Existencia de antecedentes familiares | Antecedentes familiares | Categórica |

**Proceso de Extracción**

Se extrajeron las distintas variables mencionadas desde los orígenes de EPS y póliza. Para ello la herramienta utilizada fue Teradata y el lenguaje, SQL. Adicionalmente se extrajeron algunas variables por medio de expresiones regulares sobre historia clínica.

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre Variable | Grupo Variable | Fuente Póliza | Fuente EPS |
| Asegurado ID | Datos de Identificación | V\_CLIENTE | VM\_EPS\_AFILIADO |
| Afiliado ID | Datos de Identificación | V\_CLIENTE | VM\_EPS\_AFILIADO |
| DNI Cliente | Datos de Identificación | V\_CLIENTE | VM\_EPS\_AFILIADO |
| Indicador de Negocio | Datos de Identificación | V\_CLIENTE | VM\_EPS\_AFILIADO |
|  |  |  |  |
| Edad | Sociodemográficas | V\_CLIENTE | VM\_EPS\_AFILIADO |
| Raza | Sociodemográficas |  | V\_RAZA  VM\_IPS\_HC\_AFILIADO |
|  |  |  |  |
| Cigarrillo | Hábitos |  | VS\_PEP\_HABITOS\_HIST\_POB |
| Alcohol | Hábitos |  | VS\_PEP\_HABITOS\_HIST\_POB |
| Ejercicio | Hábitos |  | VS\_PEP\_HABITOS\_HIST\_POB |
|  |  |  |  |
| IMC | Resultados clínicos | MEGADATA | EVE\_PEP\_MEDICION\_RESULTADOS\_HIST |
| Obesidad | Resultados clínicos | MEGADATA | EVE\_PEP\_MEDICION\_RESULTADOS\_HIST |
| Sobrepeso | Resultados clínicos | MEGADATA | EVE\_PEP\_MEDICION\_RESULTADOS\_HIST |
| Valor Progesterona, Estradiol, LH,FSH, Prolactina | Resultados clínicos |  | VE\_IPS\_HC\_EVE\_RESU\_AYU\_DX |
|  |  |  |  |
| Hiperplasia Ductal | Antecedentes Personales |  | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE |
| Ooforectomía Bilateral | Antecedentes Personales | MDB\_CONSULTAS.VS\_SINIESTRO\_SAL | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE  VE\_EPS\_HIST\_AUT\_PRESTACION |
| Gen BRC | Antecedentes Personales |  | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE  VE\_EPS\_HIST\_AUT\_PRESTACION |
| Terapia Hormonal | Antecedentes Personales | MDB\_CONSULTAS.VS\_SINIESTRO\_SAL | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE  VE\_IPS\_EVE\_GINECOVASCULAR |
| Edad Menopausia | Antecedentes Personales |  | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE  VE\_IPS\_EVE\_GINECOBSTET |
| Edad Menarca | Antecedentes Personales |  | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE  VE\_IPS\_EVE\_GINECOBSTET |
| Edad Primer parto | Antecedentes Personales |  | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE  VE\_IPS\_EVE\_GINECOBSTET  VM\_EPS\_MIG\_BMH\_ID\_EMBARAZO |
| Indicador Embarazo | Antecedentes Personales | V\_CLIENTE | VE\_IPS\_HC\_EVE\_MOT\_CLIENTE  VE\_IPS\_EVE\_GINECOBSTET  VM\_EPS\_MIG\_BMH\_ID\_EMBARAZO |
| Indicador BIRADS | Antecedentes Personales |  | VE\_IPS\_HC\_EVE\_RESU\_AYU\_DX |
| Indicador Biopsia mama | Antecedentes Personales |  | VE\_IPS\_HC\_EVENTO\_PROCED |

Siguiendo la metodología de índice de desenlaces, se crearon tablas temporales donde se tuvieran las variables anteriormente descritas a nivel de DNI\_Cliente de la población. Se construyó el query del archivo **cm.sql**, donde se realizaron los joins entre la base poblacional y las variables extraídas.

La fase de modelación se ejecuta en la herramienta Jupyter Notebook, con lenguaje Python. Ésta inicia con la ejecución del archivo templates/summary\_analysis.cm, donde se realiza la lectura del query cm.sql y la conversión de la información a formato parquet, para ello se escribe en la carpeta data el archivo data\_cm.parquet.

Posteriormente se ejecuta el notebook output/summary\_analysis\_cm, allí se realizan dos tipos de análisis descriptivos sobre cada una de las variables obtenidas:

1. **Análisis incondicional:** consta de la obtención de estadísticos sobre la variable. Para el caso de variables numéricas, como la edad, se obtienen descriptivos como la media, la varianza, los cuartiles, un boxplot, entre otros. Para variables categóricas, como la raza, se tiene un conteo de frecuencias sobre cada categoría de la variable.
2. **Análisis condicional:** consta del análisis incondicional pero realizado sobre cada categoría de la variable objetivo. Para el presente caso, se realiza la obtención de estadísticos en el caso de la variable Ind\_Cancer\_Mama en ‘Si’ y en ‘No’.

Si por medio del análisis descriptivo se concluye que algunas variables pueden descartarse para los pasos posteriores, éstas se introducen en una lista y al final del proceso se excluyen, sobreescribiendo el archivo data\_cm.parquet.

A continuación, se tiene el notebook de modelación, cuyo input es el archivo data\_cm.parquet. Las fases de la modelación son las siguientes:

1. **Transformación de variables**

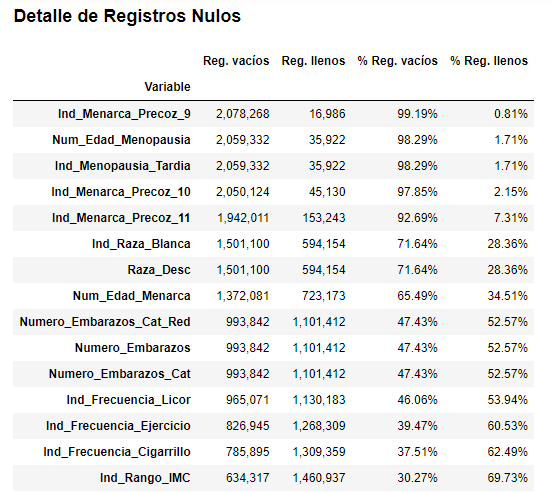
Se eliminan valores con caracteres como ‘\*’ ó ‘>’ de las variables hormonales que vienen de ayudas diagnósticas. Se realiza la conversión de éstas variables a numérico y se crean variables adicionales de tipo categórico con base en valores establecidos:

* Ind\_Progesterona\_Alta: si Valor\_Progesterona es mayor a 5 ng/dl
* Ind\_Prolactina\_Alta: si Valor\_Prolactina es mayor a 20 ng/dl
* Ind\_Estradiol\_Alta: si Valor\_Estradiol > 75 ng/dl
* Ind\_FSH\_Alta: si Valor\_FSH>10 ng/dl
* Ind\_LH\_Alta: si Valor\_LH > 10 ng/dl
* Ind\_Bajo\_Peso: si IMC < 20 kg/m2
* Ind\_SobrePeso: si IMC > 25 kg/m2
* Ind\_Obesidad: si IMC > 20 kg/m2
* Ind\_Birads: si por lo menos alguna de las variables Ind\_Birads4, 4A, 4B, 4C ó 5 es igual a 1.

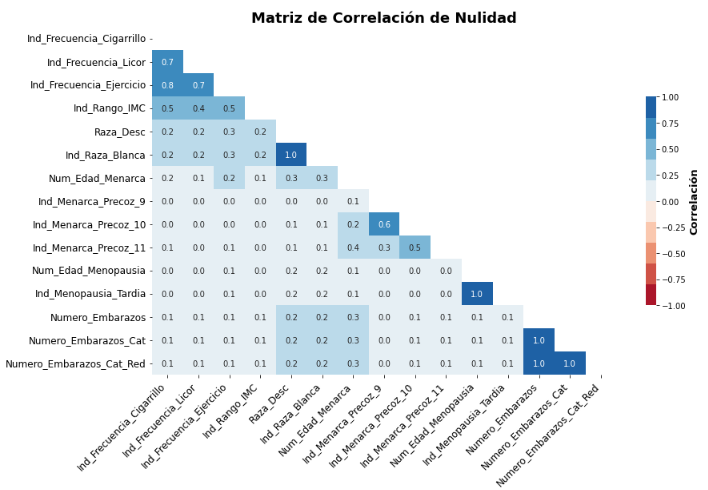
1. **Imputación de variables**

Se realiza la imputación sobre los registros nulos de algunas variables para conservar la mayor cantidad de registros posibles. La técnica usual es imputar las variables categóricas con el valor más común y las numéricas con la moda o el promedio.

* Ind\_Menarca\_Precoz\_12: se imputa con ‘N’
* Ind\_Embarazo: se imputa con ‘N’
* Valor\_Progesterona: se imputa con la moda
* Valor\_Prolactina: se imputa con la moda
* Valor\_Estradiol: se imputa con la moda
* Valor\_LH: se imputa con la moda
* Valor\_FSH: se imputa con la moda
* IMC: se imputa con el promedio

1. **Datos faltantes**   
   Se evalúa el porcentaje de registros llenos para cada una de las variables obtenidas. Los resultados obtenidos se observan a continuación:  
   

Posteriormente, se obtiene la matriz de correlación de nulidad, donde los coeficientes de correlación superiores a 0.5 sugieren la eliminación de variables para evitar la multicolinealidad:



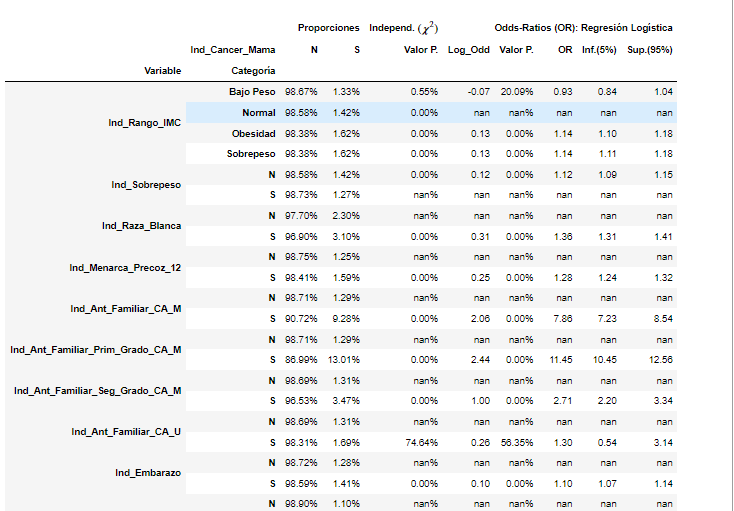
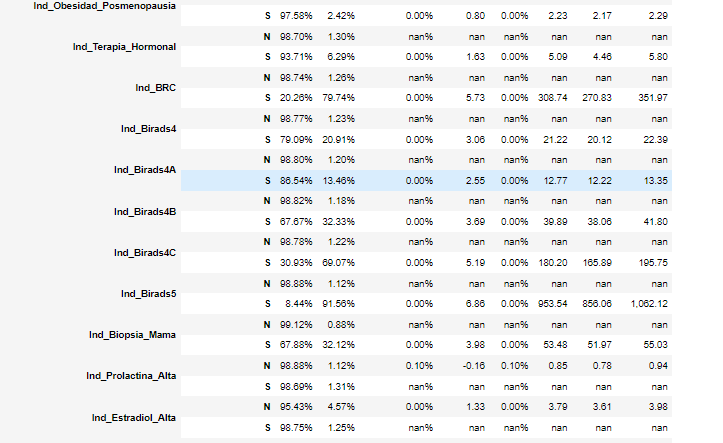
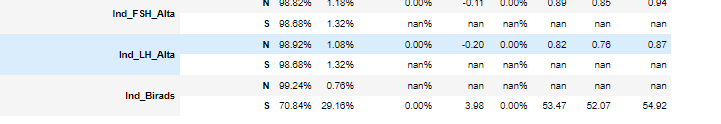
Con base en los resultados obtenidos, se considera la eliminación de las siguientes variables, tanto por completitud de registros como por multicolinealidad:

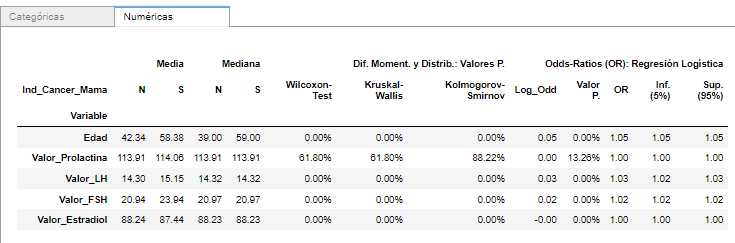
* Ind\_Menarca\_Precoz\_9
* Ind\_Menarca\_Precoz\_10
* Ind\_Menarca\_Precoz\_11
* Numero\_Embarazos\_Cat
* Numero\_Embarazos\_Cat\_Red
* Ind\_Menopausia
* Ind\_Biopsia\_Hiperpl\_Ductal
* Raza\_Desc
* IMC
* Numero\_Embarazos
* Num\_Edad\_Menopausia
* Num\_Edad\_Menarca
* Ind\_Menopausia\_Tardia
* Ind\_Frecuencia\_Cigarrillo
* Ind\_Frecuencia\_Ejercicio
* Ind\_Frecuencia\_Licor
* Ind\_Obesidad
* Ind\_Ooforectomia\_Bilateral

Finalmente, con el objetivo de mejorar la calidad de los datos, se eliminan los registros con por lo menos un registro nulo.

1. **Efectos Individuales**  
     
   Se considera el peso de cada variable sobre la variable dependiente y, con base en lo observado, se descartan variables que no sean significativas o impacten a la objetivo.

Registros completos bivariados: se utilizan pruebas de diferencias de momentos, distribuciones y regresiones logísticas entre cada una de las variables y la variable objetivo. Es de interés conocer los valores P de las pruebas de hipótesis realizadas, donde un valor inferior al 5% sugiere que la variable es suficientemente significante para entrar al análisis. Igualmente, conviene conocer el odd-ratio producto de cada una de las regresiones logísticas individuales. Teóricamente, un odd-ratio superior a 1 da indicios de que la variable tiene un impacto sobre la dependiente. Los resultados del análisis se observan a continuación:

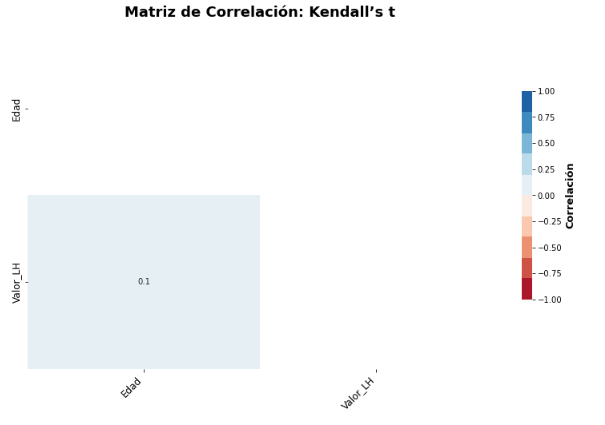
  
  




Con base en los criterios mencionados, las siguientes variables se excluyen del análisis debido a que no poseen una significancia (valor P superior al 5%) o un poder predictivo (odd-ratio < 1) suficientes:

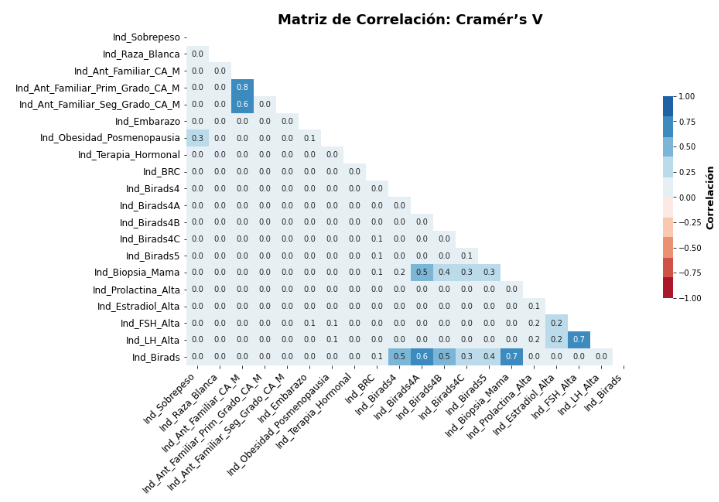
* Ind\_Ant\_Familiar\_CA\_U
* Ind\_Menarca\_Precoz\_12
* Valor\_Prolactina
* Valor\_FSH
* Valor\_Estradiol
* Ind\_Rango\_IMC

1. **Estructura de dependencia**  
   Se ejecutan una serie de análisis para determinar el nivel de relación que tienen las variables entre sí. Para las variables numéricas, se obtiene la siguiente matriz de correlación:



En consecuencia no es necesario eliminar ninguna variable.

Para las variables categóricas se obtiene la siguiente matriz:



Obsérvese que se tienen coeficientes de correlación superiores a 0.5 en los siguientes escenarios:

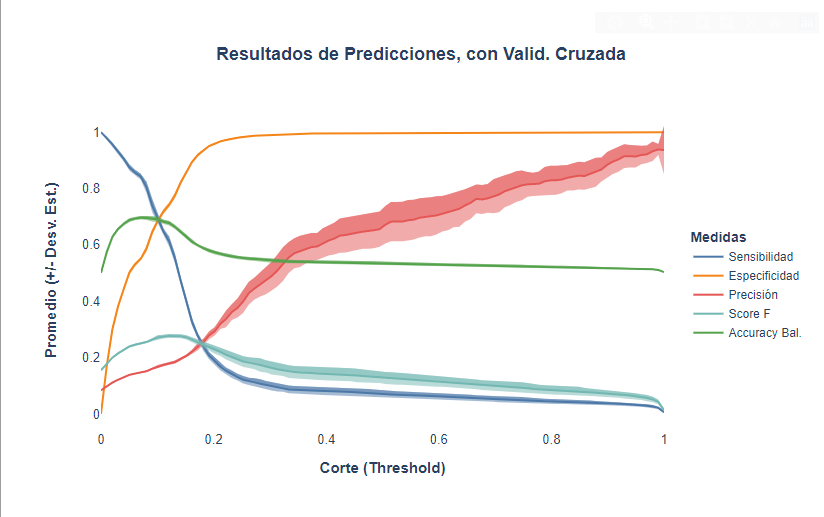


Por lo que se requiere realizar una selección entre variables correlacionadas para evitar multicolinealidad. Las variables seleccionadas son:

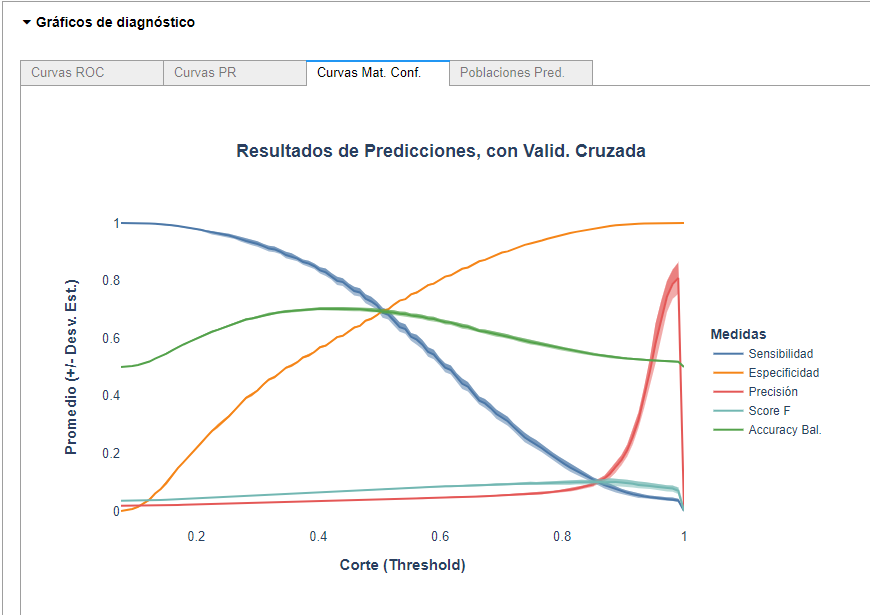
* Ind\_LH\_Alta
* Ind\_Birads
* Ind\_Ant\_Familiar\_CA\_M

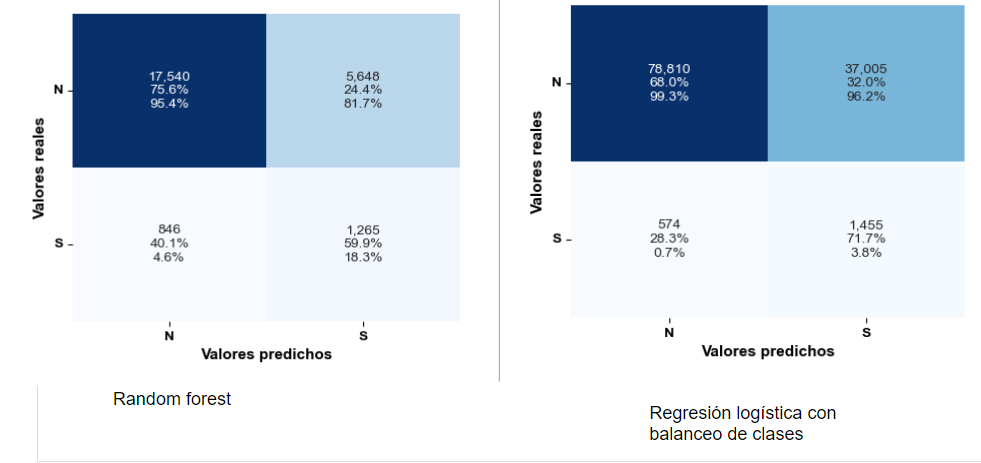
La selección de las variables Birads y antecedente familiar tiene la ventaja de poseer una mayor completitud de registro respecto a las otras variables correlacionadas: Ind\_Birads combina todos los casos de 4 a 5 e Ind\_Ant\_Familiar\_CA\_M combina los casos de primer y segundo grado.

1. **Especificación del modelo**  
     
   Se realiza la estimación del modelo y las variables finales con base en métodos de estandarización y de selección recursiva. Se compone de varias fases:
2. Datos de modelación: se estandarizan y codifican las variables de entrada.
3. Estimación: se ejecutan regresiones logísticas simples de la variable dependiente en función de cada una de las variables de entrada, con el fin de determinar el efecto de cada una de ellas.
4. Selección recursiva de variables: se determina a partir del conjunto inicial de variables y, con base en alguna métrica (precisión, recall, f-score, etc) cuáles maximizan o minimizan la métrica, dentro de un escenario de validación cruzada. Se obtiene como salida el conjunto de variables definitivo a ingresar al modelo.
5. **Modelación:**   
   Se ejecuta la partición de la data en subconjuntos de entrenamiento y validación y se determina el mejor modelo con base en la maximización o minimización de una métrica. Para el presente caso se tomó una proporción del 80-20 en entrenamiento y validación. Se determinó maximizar la precisión. Se realizaron pruebas con modelos de random forest y regresión logística con balanceo de clases, obteniendo las siguientes curvas de resultado de predicciones con validación cruzada:



Random forest

  
Regresión logística con balanceo de clases.

Y obteniendo las siguientes matrices de confusión;:  


Se determinó que la mejor familia de modelos es la **regresión logística con balanceo de clases**, debido a una minimización en los falsos negativos y un menor tiempo de entrenamiento en términos computacionales.

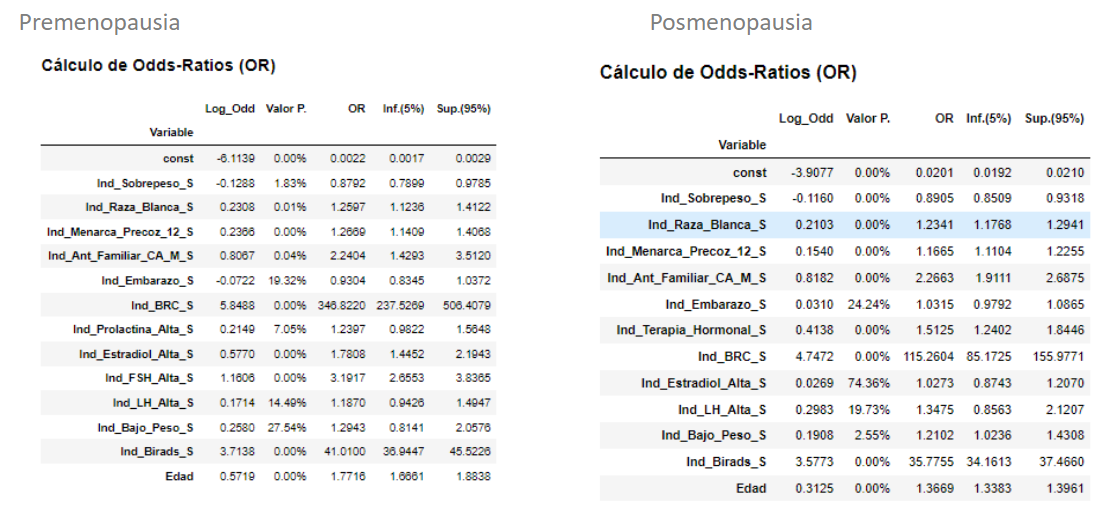
**Fase II: modelación de pre y posmenopausia**

Con base en la literatura revisada y la socialización con el negocio, se determinó realizar dos modelos: uno para población premenopáusica y otro para la población posmenopáusica. Para ello se definió la menopausia con base en los siguientes criterios:

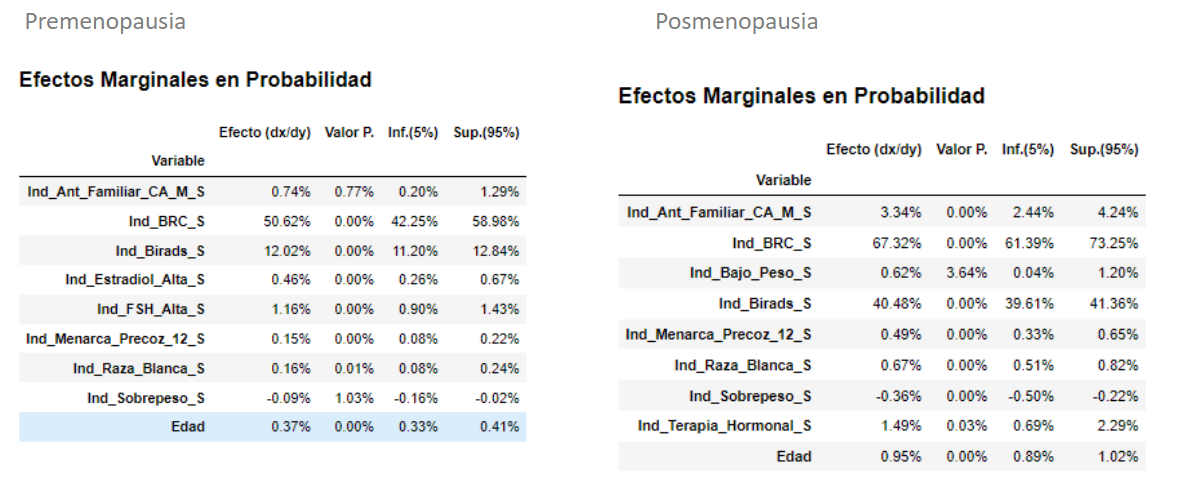
* Si en alguna de las fuentes (ginecobstétrica, historia clínica, etc) se reporta la menopausia de la paciente.
* Si la paciente ha tenido el procedimiento de ooforectomía bilateral.
* Si se desconoce el estado de menopausia de la paciente, se asume que ésta es posmenopáusica si tiene una edad mayor o igual a 48 años, correspondiente al promedio de los datos conocidos.

El proceso para la elaboración de los dos modelos se encuentra en dos notebooks: correspondientes a cada uno de los escenarios. Los pasos ejecutados para el análisis y modelación son los mismos que los descritos para el dataset completo, con la diferencia de que se excluyen del análisis las variables asociadas a la menopausia (Ind\_Menopausia, Num\_Edad\_Menopausia, Ind\_Ooforectomia\_Bilateral, Obesidad\_Posmenopausia).

Los efectos bivariados junto con los odd-ratios de los dos escenarios se observan a continuación:



Con base en los resultados anteriores se determina la exclusión de las variables [Ind\_Embarazo, Ind\_Prolactina\_Alta, Ind\_LH\_Alta e Ind\_Bajo peso] para la población premenopáusica; y la exclusión de [Ind\_Embarazo, Ind\_Estradiol\_Alta e Ind\_LH\_Alta] para la población posmenopáusica.

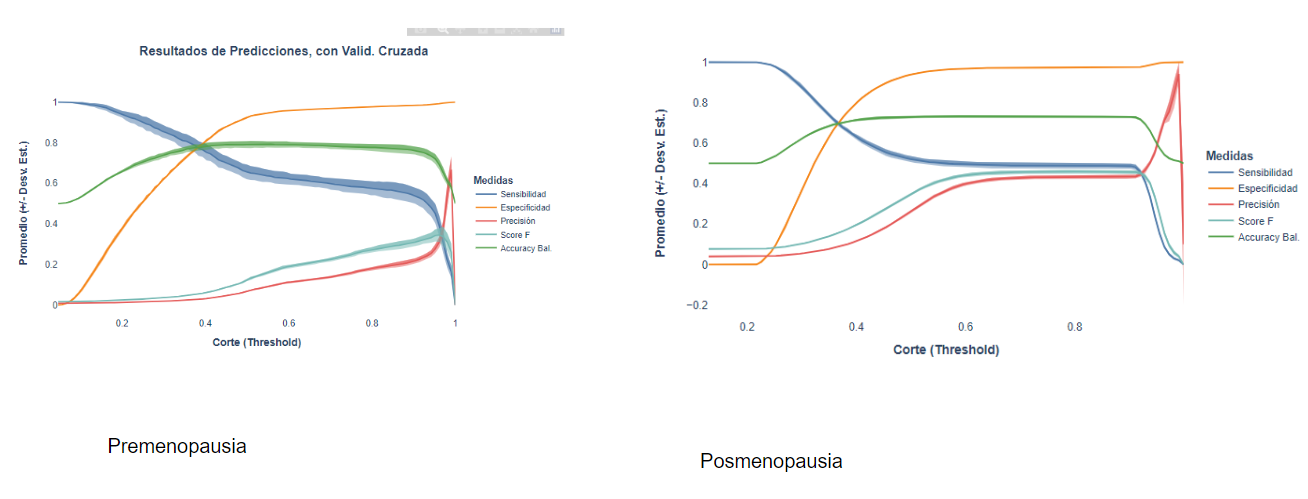
Los efectos marginales en probabilidad obtenidos se observan a continuación:  


En consecuencia, las variables de entrada para el escenario de premenopausia son:

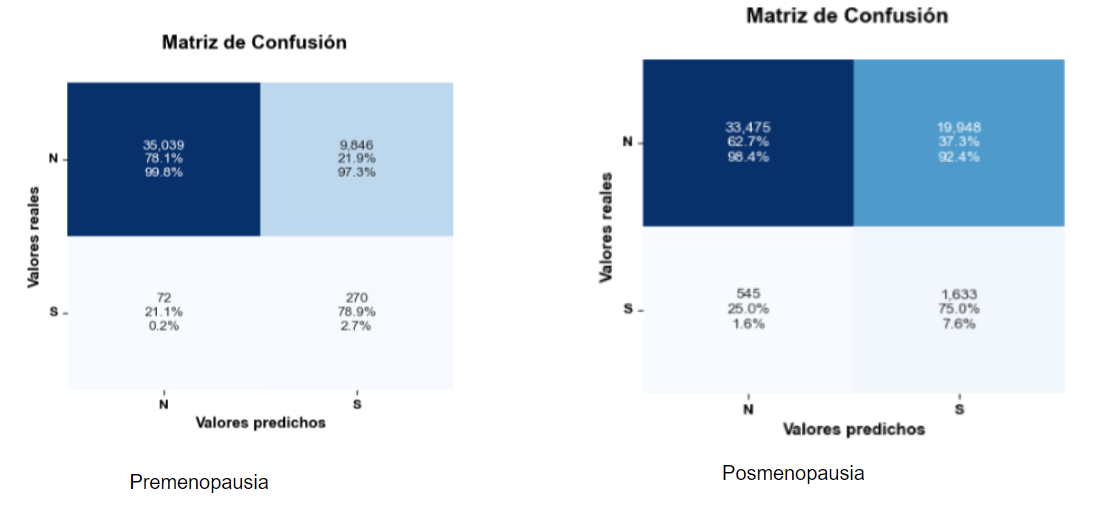
* Edad
* Ind\_Ant\_Familiar\_CA\_M
* Ind\_BRC
* Ind\_Birads
* Ind\_Estradiol\_Alta
* Ind\_Menarca\_Precoz\_12
* Ind\_Raza\_Blanca
* Ind\_Sobrepeso

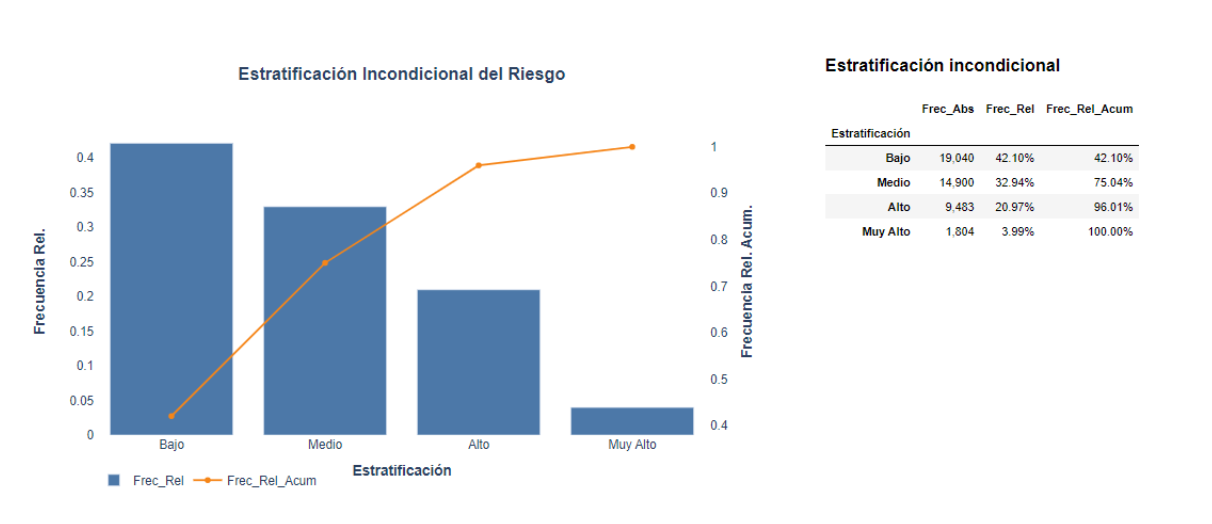
Y para el escenario de posmenopausia son:

* Edad
* Ind\_Ant\_Familiar\_CA\_M
* Ind\_BRC
* Ind\_Bajo\_Peso
* Ind\_Birads
* Ind\_Estradiol\_Alta
* Ind\_Menarca\_Precoz\_12
* Ind\_Raza\_Blanca
* Ind\_Sobrepeso
* Ind\_Terapia\_Hormonal

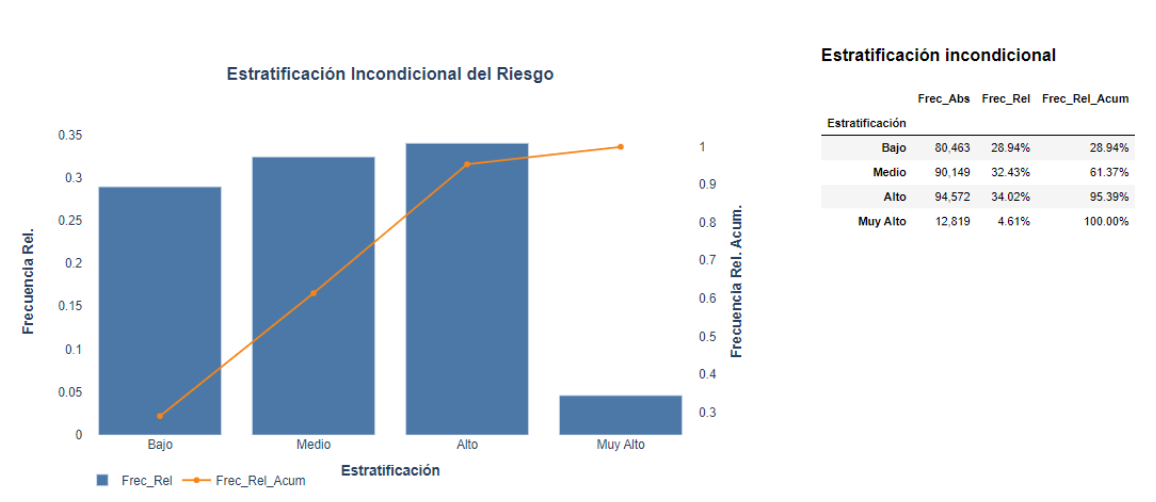
El modelo seleccionado para ambos escenarios, con base en los resultados para la población completa es la **regresión logística con balanceo de clases**, optimizada para la precisión. Las curvas de los resultados de los modelos para cada uno de los escenarios son las siguientes.  


Y las matrices de confusión asociadas:

  
Obsérvese que existe una mayor población, porcentualmente hablando, predicha en ‘S’ para el escenario de posmenopausia, respecto a premenopausia. Esto puede deberse a la influencia de la variable ‘Edad’.

La estratificación de la población que se gestionaría puede hacerse desde el riesgo (las probabilidades predichas superiores al punto de corte del modelo) o desde la salud ( las probabilidades inferiores). De los cuatro niveles de estratificación (Bajo, Medio, Alto, Muy Alto) es de interés la población en Alto y Muy alto. Para la población de premenopausia se obtuvo la siguiente estratificación:  


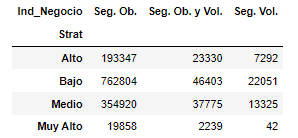
Y para la población de posmenopausia:



El empaquetamiento del modelo para un escenario de producción se resume en los siguientes objetos:

* X\_vars: consiste en el conjunto de variables de entrada.
* Encod: la codificación de las variables.
* Transf: Las transformaciones realizadas a las variables.
* Model: el modelo seleccionado en formato .pkl.
* Inv\_Score: si se realiza el análisis en función de la salud (True) o del riesgo (False).
* Strat: los criterios de estratificación de la población.
* Imput: las reglas de imputación para los valores nulos de cada variable.

Se realizó la carga de la población femenina activa a la fecha del 2022-08-02 y se excluyó la que tuviese la marca de cáncer de mama, con el fin de obtener a la población sana. Se extrajeron las variables de entrada y se realizó la predicción del desenlace. La estratificación obtenida a nivel de negocio, para la población premenopáusica, se observa a continuación:



Y para la población posmenopáusica:

